PROBLÉM KONKURUJÍCÍCH SI RIZIK A JEJICH IDENTIFIKACE

ON PROBLEM OF COMPETING RISKS AND THEIR IDENTIFICATION

Petr Volf

Adresa: Institute of Information Theory and Automation AS CR,
Pod Vodárenskou věží 4, 182 08, Prague 8

E-mail: volf@utia.cas.cz

Abstrakt: Příspěvek je věnován problému konkurujících si rizik ve statistické analýze přežití. Komplikace v analýze těchto případů je způsobena tím, že příslušné náhodné veličiny mohou být závislé. Nejprve ukážeme, jak je možné konzistentně odhadnout t.zv. incidenční funkce. Dále se zabýváme vztahy mezi marginálními, simultánním a incidenčními distribucemi v případě, že je simultánní rozdělení vyjádřeno pomocí kopuly.

Klíčová slova: Analýza přežití, konkurující si rizika, incidence, kopula.

Abstract: The contribution deals with the problem of competing risks (of competing events) in the statistical survival analysis. The case is complicated by the fact that the potential occurrence of both events may be dependent. We recall the notion of incidence function and the methods of statistical incidence analysis. Then we study the relationship between marginal, joint and incidence distributions of events when the joint distribution is modeled via a copula.

Keywords: Survival analysis, competing risks, incidence, copula.

1. Competing risks problem

Let us consider random times to certain competing (two or more) events, for instance a failure of a device caused by one of several possible causes. An underlying model assumes that there are K possibly dependent random variables T_j, $j = 1, \dots, K$. Typically, we also have to add a censoring variable C, independent of all T_j's. It is further assumed that observation of the object (device) ends with the first occurring event (or by censoring). Hence, we observe $Z = \min(T_1, \dots, T_K, C)$ and we know also what was the cause, so that we observe an indicator $\delta = 1, \dots, K, 0$ if $Z = T_1, \dots, T_K, C$, respectively.

It is known (e.g. Tsiatis, 1975) that, in general, from such observations (N i.i.d. couples (Z_i, δ_i), $i = 1, \dots, N$) it is not possible to identify neither the joint distribution of (T_j) nor their marginal distributions. On the other hand, the data allow to estimate consistently joint distributions of $(T_j, \delta = j)$ and corresponding sub-distribution functions called the (cumulative) incidence functions.

The situation can be better when our information is richer thanks to dependence of data on observed covariates. We shall recall briefly an identifiability results of Heckman and Honoré [2]. Part 3 then deals with a copula model applied to the competing risks case. Finally, an example shows the use of Gauss copula and estimation of incidence functions.

2. Incidence function

The structure of data observed in the competing risks setting enables us to estimate, consistently, the following characteristics: First, the distribution of $Z = \min(T_1, \dots, T_K)$, namely $S(t) = P(Z > t) = P(T_1 > t, \dots, T_K > t) = \overline{F}_K(t, \dots, t)$, where by $\overline{F}_K(t_1, \dots, t_k)$ we denote the joint survival function of $T_1, \dots, T_K$. Further, we can estimate the **incidence densities**

$$f_j^*(t) = P'(Z = t, \delta = j) = -\frac{\partial \overline{F}_K(t_1, \dots, t_K)}{\partial t_j}\bigg|(t_1 = \dots = t_K = t),$$

and their integrals, **cumulative incidence functions** $F_j^*(t) = \int_0^t f_j^*(s)\,\mathrm{d}s = P(Z \le t, \delta = j)$. Notice that $\lim F_j^*(t) = P(\delta = j) < 1$ if $t \to \infty$ and $S(t) = 1 - \sum_{j=1}^K F_j^*(t)$.

Another (equivalent, however more practical for estimation) definition of the cumulative incidence function is based on the cause-specific hazard functions for events $j = 1, 2, \dots, K$,

$$h_j^*(t) = \lim_{d \to 0} \frac{P(t \le Z < t + d,\, \delta = j \,|\, Z \ge t)}{d}.$$

Overall hazard rate for $Z = \min(T_1, \dots, T_K)$ is then

$$h(t) = \lim_{d \to 0} \frac{P(t \le Z < t + d \,|\, Z \ge t)}{d} = \sum_{j=1}^K h_j^*(t),$$

corresponding integrals are cumulated hazard rates $H_j^*(t)$, $H(t)$. Finally, the overall survival function $S(t) = P(Z > t) = \exp(-H(t))$. Then $f_j^*(t) = h_j^*(t) \cdot S(t)$ and cumulative incidence functions can be written as

$$F_j^*(t) = P(Z \le t, \delta = j) = \int_0^t S(s) \cdot h_j^*(s)\,\mathrm{d}s.$$

2.1. Estimation method

Let us here recall some standard notation. $N_{ij}(t)$ is the counting process with value 0 at $t = 0$ and with step $+1$ at the moment when event of type j is observed on object i. Further, let $Y(t)$ denote the number of objects in the risk set at (just before) time t, i.e. of objects without any event and not censored before t. All cumulative hazard rates can be estimated standardly by the Nelson-Aalen estimator, namely

$$\widehat{H}_j^*(t) = \int_0^t \sum_{i=1}^n \frac{\mathrm{d}N_{ij}(s)}{Y(s)}, \qquad \widehat{H}(t) = \sum_{j=1}^K \widehat{H}_j^*(t). \tag{1}$$

Overall survival function can then be estimated by the Kaplan Meier "Product Limit" (PL) estimator, or by $\widehat{S}(t) = \exp(-\widehat{H}(t))$. Asymptotic properties of estimates of incidence functions

$$\widehat{F}_j^*(t) = \int_0^t \widehat{S}(s)\,\mathrm{d}\widehat{H}_j^*(s) \tag{2}$$

follow from good asymptotic properties of $\widehat{S}$ and $\widehat{H}_j^*$ and are derived for instance in Lin [3]. In general, limit distribution of $\sqrt{n}(\widehat{F}_j^*(t) - F_j^*(t))$ is that of Gauss random process, with estimable covariance structure. As it is not a martingale, further inference (e.g. statistical tests) is not easy. Notice, however, that in the simplest case without censoring $F_j^*(t)$ and $S(t)$ correspond, at each fixed t, to probabilities in a multinomial distribution, the estimates correspond to relative occurrence, so that their properties simplify. In general, confidence regions for statistical testing are obtained by a Monte Carlo random generation.

2.2. Non-identifiability

A. Tsiatis [5] has shown that for arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish the models. Namely, this "independent" model is given by cause-specific hazard functions $h_j^*(t)$. In a parametric setting it also means that even if the MLE yields consistent estimates, we don't know parameters of which multivariate model are estimated.

On the other hand, Heckman and Honoré [2], and then others, have proved, under suitable conditions, that in the case of regression models (they considered Cox or AFT models), when our information is enriched due to knowledge of covariate values, the competing risk data suffices for full identification of the model.

3. Competing risk and copula

In the sequel we shall consider just a couple of competing events, $K = 2$, represented by random variables S, T. From the above it follows that without some knowledge about mutual dependence of S, T we are not able, in general, to estimate their distribution. The copulas offer a possibility how to model two (or multi-) dimensional distributions. Let us recall that Sklar's theorem ensures that to each 2-dimensional distribution function (of a continuous-type distribution) there exists an unique function $C(u, v)$, a distribution function on $(0, 1)^2$, such that

$$F_2(s, t) = C(F_S(s),\, F_T(t)), \tag{3}$$

where F_S, F_T are marginal distribution functions of variables S, T. The marginals of $C(u, v)$ correspond to random variables $U = F_S(S),\ V = F_T(T)$ and have uniform distribution on $(0, 1)$. There are several classes of copulas analyzed theoretically or used practically (cf. Cherubini et al. [1]). Zheng and Klein [6] showed that in the competing risks setting, when the copula function is given (assumed), marginal distributions of S, T, and then also joint distribution from (3), are estimable. They also proposed a procedure of the non-parametric estimation, proved asymptotic results and showed that their estimator reduces to the Kaplan-Meier PL estimator if S, T are independent.

3.1. Use of Gauss copula

Let X, Y be standard normal random variables $N(0, 1)$ tied with (Pearson) correlation $\rho = \rho(X, Y)$. We denote ϕ, φ univariate standard normal distribution function and density and by $\phi_2(x, y)$, $\varphi_2(x, y)$ corresponding 2-dimensional functions. Then

$$\varphi_2(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\Big\} \tag{4}$$

with $\boldsymbol{x} = (x, y)'$ and Σ the 2×2 covariance matrix with rows $(1, \rho)$ and $(\rho, 1)$. If we define $U = \phi(X),\ V = \phi(Y)$, we obtain a 2-dimensional distribution on $(0, 1)^2$ with the copula

$$C(u, v) = \phi_2\big(\phi^{-1}(u),\, \phi^{-1}(v)\big). \tag{5}$$

Naturally, $\rho(U, V) \neq \rho(X, Y)$ (though they are rather close, as a rule), while Spearman's correlations coincide, namely $\rho_{\mathrm{SP}}(X, Y) = \rho_{\mathrm{SP}}(U, V) = \rho(U, V)$.

We are, however, primarily interested in the model for dependence of competing variables S, T. Let us assume that their joint distribution function fulfils (3), where $C(u,v)$ is the copula (5). It further follows that

$$F_2(s,t) = \phi_2\big(\phi^{-1}(F_S(s)),\, \phi^{-1}(F_T(t))\big), \tag{6}$$

and, inversely, $S = F_S^{-1}(\phi(X))$, $T = F_T^{-1}(\phi(Y))$. Hence, again $\rho_{\mathrm{SP}}(S,T) = \rho_{\mathrm{SP}}(U,V)$, and "initial" $\rho = \rho(X,Y)$ is the only parameter describing the dependence of S and T. It, naturally, differs from $\rho(S,T)$, however, all values $\rho(S,T)$ (at least from $(-1,1]$) can be achieved by convenient choice of $\rho(X,Y)$. Such a flexibility is not common to many other copula types. Let us remark here that the real dependence among S,T can be much more complicated, nevertheless the use of Gauss copula model offers rather simple and sufficiently flexible (as regards the correlation) set of distributions.

3.2. Estimation in model with Gauss copula

When parameter ρ is known, copula (5) is fully defined and from Zheng, Klein [6] it follows that the distribution of (S,T) can be estimated, in parametric and even non-parametric setting. On the other hand, when marginal distributions F_S, F_T are known and both (3) and (5) hold with the same copula, then $\rho = \rho(X,Y)$ is estimable, and then also is the joint distribution $F_2(s,t)$. The estimation procedure is based on the maximum likelihood method. The data are (Z_i, δ_i), $i = 1,\dots,N$. The likelihood function then has the form

$$L = \prod_{i=1}^{N} \left\{-\frac{\partial}{\partial s}\overline{F}_2(s,t)\right\}^{I[\delta_i=1]} \times \left\{-\frac{\partial}{\partial t}\overline{F}_2(s,t)\right\}^{I[\delta_i=2]} \times \overline{F}_2(s,t)^{I[\delta_i=0]},$$

evaluated at $s = t = Z_i$, with $\overline{F}_2(s,t) = P(S > s, T > t) = 1 - F_S(s) - F_T(t) + F_2(s,t)$. It is due transformation (3) and (5) that $F_2(s,t) = \phi_2(x,y)$ with $x = \phi^{-1}(F_S(s))$, $y = \phi^{-1}(F_T(t))$. Hence, when we put $X_i = \phi^{-1}(F_S(Z_i))$, $Y_i = \phi^{-1}(F_T(Z_i))$, we obtain after some computation – integration of 2-dimensional Gauss density $\varphi_2(x,y)$, that

$$\begin{aligned} L &= \prod_{i=1}^{N} \left\{f_S(Z_i)\left[1-\phi_1(Y_i;\rho X_i, 1-\rho^2)\right]\right\}^{I[\delta_i=1]} \times \\ &\times \left\{f_T(Z_i)\left[1-\phi_1(X_i;\rho Y_i, 1-\rho^2)\right]\right\}^{I[\delta_i=2]} \times \\ &\times \{1 - F_S(Z_i) - F_T(Z_i) + \phi_2(X_i, Y_i\}^{I[\delta_i=0]}, \end{aligned}$$

where $\phi_1(x;\mu,\sigma^2)$ denotes the distribution function of normal distribution $N(\mu,\sigma^2)$, evaluated at x. It is seen that the problem of maximization has to be solved by a convenient search procedure. Parameter ρ is hidden in ϕ_1 and in ϕ_2. Distributions of S and T are present both explicitly and also implicitly, in transformed X_i, Y_i. Nevertheless, experience suggests that solution of both problems (estimate of F_S, F_T for given ρ, estimate of ρ for given F_S, F_T) are solvable and have unique solution.

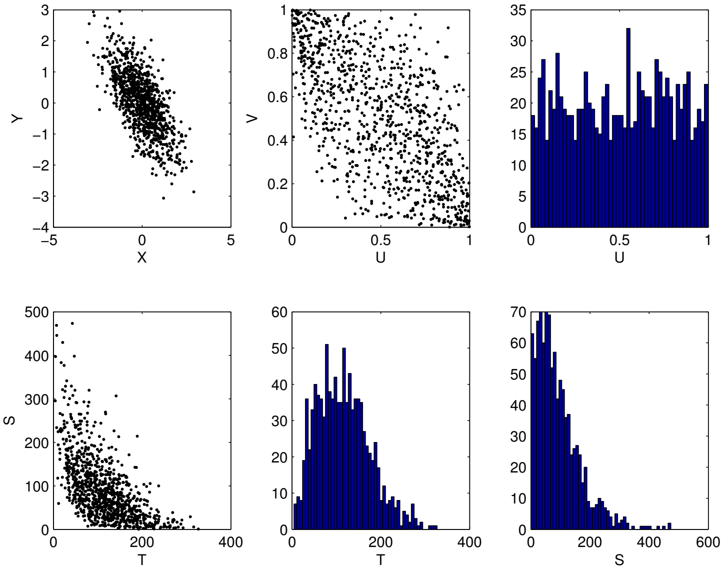

Figure 1: Scatter-plots and histograms of generated representation of X, Y, then transformed to U, V and S, T, the case with $\rho = -0.7$, N=1000.

4. Example using Gauss copula

We fixed both competing risks distributions, namely $S \sim$ Weibull ($a_s = 100$, $b_s = 1.2$), $T \sim$ Weibull ($a_t = 130$, $b_t = 3$), and censoring variable $C \sim |\text{Normal}(\mu = 150,\ \sigma = 50)|$. The rate of censoring was among 10 − 20 %. Weibull distribution function was taken in form $F(s) = 1 - \exp\left(-\left(\frac{s}{a}\right)^b\right)$, $s > 0$. The analysis was done for two values of ρ, namely for $\rho = 0.5$ and $\rho = -0.7$.

First, we show how the data (X, Y) generated from ϕ_2 are transformed to (U, V) by (5) and then to (S, T) by (3). Figure 1 shows the scatter-plots and

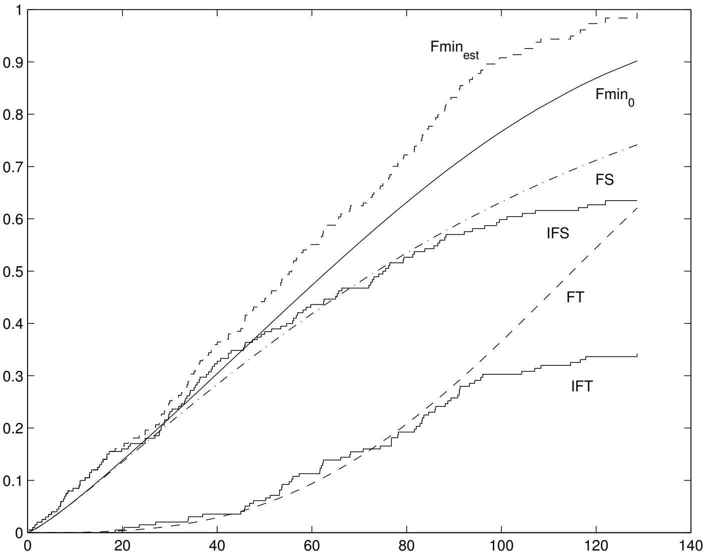

Figure 2: "True" distribution functions F_S, F_T, $F\min_0$ under independence hypothesis, estimated $F\min_{\text{est}}$, estimated cumulative incidence functions IF_S, IF_T. Case of $\rho = -0.7$, $N = 200$.

histograms of generated values ($N = 1000$), in the case $\rho = -0.7$. We also computed distributions numerically. Numerically computed correlations yield $\rho(U, V) = 0.432$, $\rho(S, T) = 0.376$ in the case with $\rho = 0.5$, and $\rho(U, V) = -0.685$, $\rho(S, T) = -0.625$ in the case with $\rho(X, Y) = -0.7$.

Further, we show an example of estimates of cumulative incidence functions, following the approach described in part 2. The same type of data as in previous example was generated. We display here just the case of $\rho = -0.7$, $N = 200$. Figure 2 shows both underlying "true" distribution functions F_S and F_T, and also $F\min_0$ of $\min(S, T)$ under hypothesis of independence. Dashed step-wise curve is the PL-estimate of true distribution $F\min_{\text{est}}$ of $\min(S, T)$. It differs from $F\min_0$, it could be taken as an evidence that independence hypothesis does not hold. Finally, two full step-wise curves are estimated cumulative incidence functions IF_S, IF_T of S, T, respectively. Notice that they summarize to $F\min_{\text{est}}$. Easy generation of artificial data is another advantage of Gauss copula.

In a particular case when marginal distribution are known, the hypothesis of independence (i.e. that $F\min = F\min_0$) can be tested easily with the aid of asymptotic properties of the PLE $F\,\mathrm{min}_{\mathrm{est}}$. If, moreover, the type of copula is assumed (as in our case here), parameter ρ can be estimated by the ML method and test of hypothesis on its value can be based on asymptotic normality of the MLE.

True cumulative incidence functions can be obtained by integration of expressions corresponding to the 1st and 2nd part of the likelihood function. Namely, we used numerical integration of

$$\mathrm{d}IF_S(t) = f_S(t)\left[1 - \phi_1(y; \rho x, 1-\rho^2)\right],$$
$$\mathrm{d}IF_T(t) = f_T(t)\left[1 - \phi_1(x; \rho y, 1-\rho^2)\right],$$

where again $x = \phi^{-1}(F_S(t))$, $y = \phi^{-1}(F_T(t))$.

5. Conclusion

The problem of competing risks has been studied and the difference between marginal distributions and observed incidence of events has been analyzed. The main goal was to describe the procedure of estimation of incidence functions and, further, to study the use of Gauss copula in modeling and random generation of competing risks data.

Acknowledgement

This work was supported by the GA CR grant No. 13-14445S.

References

[1] Cherubini U., Luciano E., and Vecchiato W.: *Copula Methods in Finance.* Wiley, Chichester, 2004.

[2] Heckman J. J., Honoré B. E.: The identifiability of the competing risks model, *Biometrika* **76**, 325–330, 1989.

[3] Lin D. Y.: Non-parametric inference for cumulative incidence fumctions in competing risks studies, *Statistics in Medicine* **16**, 901–910, 1997.

[4] Scheike T. H., Zhang M.: Flexible competing risks regression modelling and goodness-of-fit, *Lifetime Data Analysis* **14**, 464–483, 2008.

[5] Tsiatis A.: A nonidentifiability aspects of the problem of competing risks, *Proc. Nat. Acad. Sci. USA* **72**, 20–22, 1975.

[6] Zheng M., Klein J. P.: Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82**, 127–138, 1995.